(This is a pre-print text)

BELIEF AND MORAL RESPONSIBILITY*

Carlos J. Moya. University of Valencia (Carlos.Moya@uv.es)

Compatibilists and incompatibilists agree that a condition of control over one's choices and actions has to be satisfied if one is to be morally responsible for these choices and actions. An agent's control over her choices and actions should include at least two aspects: some degree of rationality and some degree of autonomy or self-determination. The rationale for including the first aspect is that purely random, arbitrary choices or actions do not seem to be sufficiently in the agent's hands (under her control) for her to be justifiably judged as blame- or praiseworthy on account of them. This is why they should meet at least some minimal rationality requirements. Satisfaction of the second aspect, in turn, is designed to ensure that the agent is actually the source or author of the choice or action for which *she* is judged as blame- or praiseworthy. Actions or choices caused by external forces are not justifiably attributed to an agent. This much can be considered as common ground. Any acceptable theory of moral responsibility should include some requirement of rational control and some requirement of autonomy or self-rule. Discrepancies, however, start shortly after these minimal points of agreement. Typically, incompatibilists tend to consider compatibilist accounts of the self-determination aspect as not deep enough to ground moral responsibility, while compatibilists tend to view incompatibilist theories, with their insistence on indeterminism as necessary for moral responsibility, as unable to offer a satisfactory account of the rationality aspect.

For early compatibilists, in rough terms, an agent exercises control over her actions on the basis of her desires: an agent controls her actions provided that she does what she wants to do. This view of control does include the two aspects we have pointed to: on the one hand, a choice or action caused by desire has a minimal rational explanation; on the other hand, if an agent's self – as Hume suggested – is partly constituted by her own desires, determination by her own desires is self-determination. However, control, so conceived, is arguably too superficial as a basis of moral responsibility attributions.[1] It has been replaced by more sophisticated compatibilist accounts, which incorporate deeper levels of control. The desire on which an agent acts has to be backed by a reflective,

second-order volition (Frankfurt), or by the agent's values (Watson), and the agent's self is conceived as centrally constituted by her second-order volitions or by her values, rather than by her ordinary, first-order desires. On other perspectives, the agent has to be able to form objectively correct values and to act on them (Wolf), or to act on a practical reasoning 'mechanism' that is her own and appropriately responsive to reasons (Fischer and Ravizza).

Incompatibilists may readily acknowledge a progress in compatibilist accounts of control, from the early, simple proposals of Hobbes or Hume to the sophisticated views of Frankfurt, Watson, Wolf or Fischer. But, in their opinion, owing to the project of giving an account of moral responsibility compatible with determinism, even sophisticated views are bound to offer a weak, superficial picture of control, which cannot provide an appropriate ground for moral responsibility. Incompatibilists may agree that rational and volitional control over one's actions, in some of the ways proposed by compatibilists, may be necessary for moral responsibility, but they will insist that an additional feature is required in order to have something close to a sufficient condition, namely ultimacy. In order for an agent to be morally responsible for her actions, she has to be their true, ultimate origin by having ultimate control over them. An agent enjoys rational and volitional control over her actions by choosing to perform them in the light of such factors as her first- and second-order desires, values and even traits of character constitutive of her self. This means that these factors explain, as either necessary or sufficient conditions, why she acted as she did. So, an agent's control has to extend to these explanatory factors in order for her to effectively control the choices and actions that such factors help to explain and to be truly praise- or blameworthy for such choices and actions. According to the incompatibilist's intuition, an agent's rational and volitional control over her actions is too slender a basis for moral responsibility; she also has to control the self that these actions arise from. An agent is truly morally responsible for the way she acts only if she is truly responsible for the way she is. Only then can the agent be said to be the true, ultimate source or origin of her actions and objectively deserve praise or blame for them.

This deep, ultimate control condition for moral responsibility is what Robert Kane has called 'ultimate responsibility' and Galen Strawson 'true self-determination'. If this actually is a condition for moral responsibility, the sceptical suspicion easily arises that

moral responsibility is not possible. Ultimate control involves two aspects, namely ultimacy of source and rational cum volitional control. And it would seem that this condition is incompatible with either determinism or indeterminism. Determinism may allow for rational cum volitional control, but not for ultimacy of source, for, with the possible exception of a first, uncaused cause, there are no ultimate sources or origins in a deterministic world. Indeterminism, in turn, allows for events, such as choices, that, being undetermined, can play the role of fresh, ultimate origins or causes, but now it seems that these ultimate causes cannot be under the agent's rational cum volitional control. If these events, say choices, are explained by previously existent reasons, they can be rational but hardly ultimate causes; and if they are not so explained, they can be ultimate but not rationally controlled causes.

In fact, acceptance of deep, ultimate control as a requirement for moral responsibility has led some thinkers, such as Galen Strawson and Derk Pereboom (2001), to take a sceptical stance towards moral responsibility. Strawson's well-known sceptical argument goes roughly as follows. Rational actions are paradigmatic candidates to the status of free and responsible actions, if such there are. Now the way we act when we act rationally, that is, for reasons, depends on our mental constitution, or character. So, unless we are truly responsible for our mental constitution, we will not be truly responsible for our rational actions. But in order to be truly responsible for our mental constitution, we have to have *chosen* that mental constitution in a *rational* way, that is, in the light of certain principles of choice or reasons. These reasons or principles explain the choice of our mental constitution. So we cannot be truly responsible for this choice, and so for our chosen mental constitution, unless we are responsible for having such principles to begin with; and this in turn requires that we have chosen them rationally, that is, in the light of a further set of principles of choice or reasons, and so on. According to Strawson, then, true responsibility or true self-determination (ultimate control, in our terms) 'is logically impossible because it requires the actual completion of an infinite regress of choices of principles of choice' (Strawson 1986: 29). It seems, then, that no choice can be both an ultimate and a rational source of one's actions. In the end, we are bound to choose and act on the basis of factors that we cannot have rationally chosen and for which we cannot be truly responsible. Ultimate control (true self-determination, in Strawson's terms) would seem to involve a self-defeating demand for self-creation. As Randolph Clarke has

put the point, according to Strawson 'rational free action would be possible only for an agent who was *causa sui*' (Clarke 1997: 37).

If something like ultimate control is in fact both necessary for moral responsibility and impossible to attain, then moral responsibility is not possible. And this is precisely Strawson's position. Robert Kane agrees with Strawson that ultimate control ('ultimate responsibility' in his terms) is necessary for moral responsibility in the deep sense of true desert: only if an agent is the ultimate source of her actions can it be justified to consider her as truly, objectively praise- or blameworthy for them. However, unlike Strawson, he thinks that ultimate control can be attained.[2] According to Kane, an agent can choose rationally and voluntarily her own character and motives and so be truly responsible, in Strawson's sense, for having them as well as for the choices and actions that they help to explain. Strawson's challenge immediately arises: how is it possible for an agent to choose her own character and motives rationally and voluntarily unless she already exists endowed with a previous character and motives? Kane is certainly aware of this difficulty. He agrees that his view of ultimate responsibility 'appears to lead to a vicious regress … The regress would stop with actions that were not explained by our characters and motives (or by anything else, for that matter), but then in what sense would be responsible for *such* actions?' (Kane 1996: 37).

In order to meet this crucial objection, Kane resorts to certain choices in a person's life through which she forms her own character and motives. Kane calls these choices 'Self-Forming Willings' (SFWs). If SFWs are to stop the regress that threatens ultimate responsibility they have to satisfy certain conditions. On the one hand, they must have no sufficient explanation in terms of the agent's pre-existing character, motives and preferences. They have to be genuinely open and undetermined, relative to the past and the natural laws, for otherwise the agent could not be their *ultimate* source. And, on the other hand, it is also crucial that the choice, whichever way it may go, remains under the agent's rational and volitional *control*. It must be a rational and motivated choice, a result of the agent's rational will. Irrational or arbitrary choices are not an appropriate foundation of moral responsibility, as compatibilists have always contended against incompatibilists.

Can SFWs satisfy these conditions? It is tempting, but wrong, to conceive of SFWs as Buridan's Ass cases, in which the agent has equal reasons for going one way or another, for in these cases, 'instead of one choice … being arbitrary relative to the prior deliberation, both would be arbitrary' (Kane 1996: 109). Rather, what the agent confronts in SFWs is *incommensurable* sets of reasons for going one way or another, such as moral reasons and reasons of self-interest, or prudential considerations and desires for an immediate pleasure. In cases like these, the agent is 'torn between conflicting internal points of view that represent *different and incommensurable visions of what they want in life* or what they want to become' (Kane 1996: 199). The agent's will is unsettled. She wants to act on one set of reasons and she also wants to act on the other. These choices are self-forming in that, by making them, the agent causally contributes to shaping her own character, motives and will, so that she can be said to be ultimately responsible for those psychological factors and so for further choices and actions that can flow from them. Incommensurability of reasons for each option is a crucial feature of SFW situations, which supposedly allows the final choice, whichever it is, to be the result of the agent's rational will. As Kane writes, 'for SFWs, each outcome is rational for different and incommensurable reasons' (1996: 178).

Are Kane's SFWs able to stop the regress that threatens the possibility of ultimate control? Though they go some way towards doing so, I do not think they go far enough. As we have seen, in an SFW the agent confronts a choice that she has to make on the basis of incommensurable reasons. But if reasons are actually incommensurable, how can the agent choose one of the alternatives rationally? A choice of this kind can be rationally made if the agent faces it with some criterion (a meta-criterion, let's say) that allows her to rank one set of reasons higher than the other. Such a meta-criterion might be, for example, a Frankfurtian second-order volition, according to which she prefers to be moved to act by, say, moral reasons rather than reasons of self-interest, or maybe a Watsonian valuational system, which ranks the former higher than the latter (or vice versa). However, if the agent confronts the choice with a meta-criterion, her will is not unsettled and the choice cannot be truly self-forming. She can have obtained this criterion through prior choices she faced with incommensurable reasons; but then the rationality problem arises again with regard to these choices. Or she just happens to have the criterion, but then she lacks ultimate control over it and the resulting choice. Moreover, remember that, in SFWs, the final choice, whatever it is, has to be the result of the agent's

rational will. But this condition will not be met if the agent confronts the choice with a meta-criterion, for then only some of the choices will be rational, namely those which accord with the criterion, but not those which conflict with it. Therefore an SFW should be made exclusively on the basis of the incommensurable sets of reasons the agent considers. And then it is hard to see how the choice could possibly be under her rational cum volitional control. The role of a meta-criterion is now played by the agent's pure decision: 'The agents will *make* one set of reasons or motives prevail over the others then and there *by deciding* … [B]oth options are wanted and the agents will settle the issue of which is wanted *more* by deciding' (Kane 1996: 133). Kane's libertarianism tends to turn into sheer *voluntarism* or *decisionism*. The agent's choice can be ultimate at the cost of her losing rational control over it. In fact, the two aspects of ultimate control, namely ultimacy of source and rational cum volitional control, seem to pull in opposite directions.

But then, if ultimate control is so seemingly impossible to attain, why insist on it as a necessary condition of moral responsibility? Why not to abide by relative, less radical, non-ultimate forms of control or self-determination? Compatibilists have in fact described a large variety of such forms. Won't this insistence on *ultimate* control pave the way for scepticism? I can think of two main considerations in favour of this condition. First, the requirement of ultimate control corresponds to the depth of moral responsibility attributions. A serious ascription of moral responsibility is directed to the agent herself, on the assumption that she is the true, ultimate origin, and not a mere instrumental or derived cause, of the action or consequence thereof for which we hold her responsible. These attributions have deep effects on our self-esteem and sense of dignity. So the desire for deep personal control over the grounds on which such attributions are made is clearly reasonable. We do not want our worth and value to depend on factors beyond our reach and control. Second, compatibilist construals of the control condition, which dispense with ultimacy, seem capable of being satisfied by agents who prima facie do not seem morally responsible, such as Brave New World citizens or agents in 'Covert Non-Constraining Control' situations (to use Kane's terms). These construals would appear to be too weak to ground moral responsibility understood as true, objective desert. So understood, moral responsibility would seem to require some form of deep, ultimate control over our choices and actions.

I want to suggest that, appearances to the contrary notwithstanding, deep, ultimate control is a requirement that can actually be met, provided that some unexamined assumptions about this condition, which can be found in many authors, and especially in Strawson and Kane, are brought to light and questioned. Consider Strawson. According to him, true responsibility for, or ultimate control over, one's actions requires, at the very end, that one has *chosen* one's mental constitution, the way 'one is, mentally speaking' (Strawson 1986: 28), as well as the principles on which such a choice is made. For Kane, in turn, ultimate responsibility rests upon Self-Forming Willings, undetermined choices by means of which agents build up their own character and motives. At the root of ultimate control over one's actions we find *acts of will* or *choices*. So both Strawson and Kane assume as a matter of course a will-centred view of ultimate control and moral responsibility. For them, an agent cannot have deep, ultimate control over her actions and be truly praise- or blameworthy for them unless she has *chosen* the springs of those actions. This assumption, in turn, seems to rest upon a more general view, namely that one can be said to control only that which one has a choice about. Only something that is subject to one's will could be said to be under one's control and so be an appropriate ground or object for moral responsibility attributions.

This assumption, together with a plausible rationality requirement for choices, leads quickly to scepticism about ultimate control. If we accept that ultimate control over one's choices and actions requires that one has chosen the springs of these choices and actions and that this choice has itself to be based on reasons, then we shall have to accept that, in order for us to have ultimate control over this choice, we should have chosen the reasons on which we made it; but this other choice, in turn, should also be based on reasons, which we should have chosen in the light of further reasons, and so on. We have started the regress of choices of principles of choice that Strawson rightly holds to be impossible to complete. Scepticism looms.

Another widespread and unexamined assumption in current conceptions of ultimate control is closely connected with the first. Consider Strawson's claim that, in order for true responsibility to be possible, one has to have chosen the very roots of one's choices, the principles on which one makes them. This claim, I would think, presupposes a deeply individualistic view of human agents as radically self-made, self-contained entities, whose moral responsibility is undermined by the influence of any factors that are external

to them and beyond the scope of their choice. This view transpires also in Kane's conception of ultimate responsibility as a species of self-creation. From this individualistic point of view, the social nature of human agents tends to appear as a potential threat to their being ultimate sources or origins of their own actions and so to their moral responsibility for them. This second assumption reinforces the sceptical suspicion about moral responsibility already raised by the first one.

On the background of these assumptions, my position can be stated as follows. At the foundational root of moral responsibility I shall not place conative phenomena, such as choices, but cognitive ones, such as beliefs. Especially important will be a subset of an agent's beliefs, namely her evaluative views about what is really worth pursuing or avoiding in life. Beliefs of this sort are plausibly taken to play a central role in explaining our morally relevant choices and acts. I recommend, then, a cognitive, rather than a conative, approach to moral responsibility. I accept, however, Strawson's and Kane's contention that deep, ultimate control is a requirement for moral responsibility. Now on the assumption we have referred to above that all control depends on choices or acts of will, acceptance of ultimate control seems to conflict with the view that moral responsibility rests on beliefs, unless one embraces some version of doxastic voluntarism, a rather implausible position, in my opinion.[3] However, although I share the widely held position that control is necessary for responsibility, I reject the no less widespread view that all control depends on the will. So, though belief is not voluntary, we can rightly be praised or blamed for our beliefs, for we can have over them a form of control that does not rest on acts of will or choices and is deep enough to support true praise- and blameworthiness attributions. As I shall try to show, it may be justified to grant someone deep control and authorship with regard to her factual or theoretical beliefs, as well as full praise- or blameworthiness for them, even if she has not chosen the reasons and principles on which she has formed them. In fact, in some cases we would plausibly withhold our praise if we discovered that she had chosen those reasons and principles. This suggests that something similar might apply to our evaluative beliefs. If it did, then, provided that these beliefs are among the basic explanatory roots of our actions, no regress of choices would need to start and scepticism about moral responsibility could be resisted. In connection with this, and with regard to the second assumption we have mentioned above, I shall try to show that even if someone's intellectual achievements are indebted to some external sources, this need not prevent us from justifiably granting her full praise- or

blameworthiness for them. Again, this suggests that something similar could be the case with regard to an agent's evaluative beliefs and the actions and choices flowing from them. Let me now defend this cognitive approach to moral responsibility. I do not have a conclusive argument to offer, but I can advance some considerations that show this position to be fairly plausible.

Let us note, first, that talk about responsibility for one's beliefs makes perfect sense and is rather common in everyday life. We do not find less natural to praise or blame people on account of their beliefs than of their choices or actions. Think for example of a racist person: under some circumstances, we may hold her as blameworthy for her racist views as for her racist behaviour. Moreover, it would be easy to find anywhere remarks like the following: 'How could you believe what she told you? Don't you know how often she lies?' As happens with actions, excuses can be expected in situations of this kind: 'Well, you know, this time she really looked sincere.' We sometimes blame people for being careless – or, alternatively, for being too demanding – in forming their beliefs. Attribution of responsibility for beliefs might be interpreted within a view of control as based on choice. On this view, control over our beliefs would be taken to be only indirect and derived from the voluntary control we have over our cognitive activity. It seems true that we sometimes praise or blame people for their beliefs on the basis of how they have conducted their inquiries. However, if this were the only way in which we could be said to control our beliefs, the threat of a vicious regress of choices could not be conjured up. But I think we also acknowledge another form of control over our beliefs. This control is neither voluntary nor merely indirect. It does not draw entirely on the voluntary control we may have over our epistemic activity. If we can show that sometimes responsibility for beliefs rests on a form of control over them not based on choices or acts of will, this might be an important step in a defence of moral responsibility against scepticism, at least of a Strawsonian variety. Let us see what a control of this kind is like, whether it actually exists and whether we justifiably acknowledge it in some of our ascriptions of praise- and blameworthiness for beliefs.

As an initial attempt to characterize the form of control we are after, let us reflect on some examples. Think first of a secondary school student who is trying to solve a problem of, say, physics. Suppose she performs the task carefully and obtains the right result. She has had control over both her cognitive activity and her belief about the solution to the

problem and is praiseworthy on both accounts. It is not by mere luck that she has obtained the right result. But note what having control means in this case. It does not relate to choices or acts of will in any important sense. The student's control consists rather in her yielding to the internal configuration and structure of the problem, the data, physical laws, mathematical rules, and so on. It is, so to speak, a passive form of control, which the student exercises precisely in respecting and being guided by what is there, in the problem itself. She chooses neither the data nor the physical laws or the mathematical rules. In fact, she would *lose* control over both the task and its result if she chose or decided about these things, and we would rightly blame her for doing so. Moreover, all those factors are external to her self or will: they come 'from outside'. It is also true that, without the help of her teachers, she could not have solved the problem. And nonetheless my intuition, which I hope will be widely shared, is that she has control over her belief about the problem's solution and truly deserves praise for this belief. So in this simple case the two assumptions I emphasized in Strawson's and Kane's construals of ultimate control as a requirement for true desert, namely that all control that backs true desert is based on choices, and that the influence of external factors undermines such a control, are simply absent, but true desert is still there.

In order to question these assumptions further and deepen our understanding of this sort of non-voluntary control, consider now more complex cases of belief and belief formation, such as great achievements in science or philosophy. They can harmlessly be considered as belief systems. As a preliminary remark, these cases show that authorship concerning her beliefs may be no less important for a person's worth and self-esteem than concerning her choices and actions. To mention only a few examples, remember the dispute between Newton and Leibniz about the invention of infinitesimal calculus or, in more recent times, the debate about the true discoverer of the virus causing AIDS. Think also of the strongly negative moral judgement that plagiarism deserves for most of us. Questions about real source or origin, and about corresponding praise- and blameworthiness, have no less significance in the cognitive field than in the practical one. Let us now come back to our main subject, namely the nature and existence of the form of non-voluntary control over our beliefs that we are after, by reflecting on a particular example of a complex and great intellectual achievement, to witness, Descartes' *Meditationes de prima philosophia*. I hope we shall agree that Descartes must be considered as the true author and source of this work and that he truly and justifiably

deserves our praise and gratitude for it. We find him praiseworthy not only for his effort and activity, but also for its result, the important ideas and beliefs contained in this outstanding philosophical opus. However, it is interesting to note how many of the factors that made this work possible have not their origin in Descartes himself. It is hard to see, for example, how the *Meditations*, which are usually considered as a new starting point, as the very beginning of modern philosophy, could have been written without the influence of medieval philosophy (cf. Gilson 1951). Other important 'external' factors include Plato and the Platonic tradition, ancient and modern scepticism and contemporary physiology, to mention only a few. And, nonetheless, this does not incline us to question Descartes' full authorship and praiseworthiness for his work. This seems to show that our judgements about authorship and responsibility for cognitive accomplishments do not fit Kane's or Strawson's conceptions of ultimate control, at least regarding their individualistic assumption. Though Descartes did not give origin to many aspects and elements of the *Meditations*, we readily consider him as the true, ultimate author and source of his work, and rightly so. This suggests that something similar might be the case in the practical field. But Descartes' example can also be used to dispute the other assumption indicated above, according to which control is constitutively related to acts of will and choices. Though surely the will, in the form of choices, is involved in the process of creation of the *Meditations*, Descartes' control over this work is not mainly based on voluntary acts or choices, but, to a large extent, in his respect for the internal requirements and structure of the subject matter itself, in his passively yielding to the relations of justification between propositions, to the internal connections between concepts, to the force or necessity of certain steps in the reasoning process, as well as to the empirical data he employs. As happened with our example of the student, if Descartes had made some or all of these aspects depend on his will, he would have had less control over his work and would have been less praiseworthy for it. So reflection on matters of authorship and responsibility in the realm of cognitive accomplishments does not validate the assumption that all control depends on the will. Our judgements in this field do not correspond to Strawson's and Kane's views of ultimate control. And, again, this suggests that something similar might be the case in the practical realm.

Our hope can be stated as follows. If, as we have tried to show by means of examples, control over, and true desert for, our beliefs need not rest on choices or acts of will and is not necessarily undermined by the influence of external factors; and if control over our

actions rests ultimately on control over our beliefs, then moral responsibility understood as true desert, as true praise- or blameworthiness, could be shown to be possible and not to fall prey to an infinite regress of choices or to sheer arbitrariness. My suggestion is, in fact, that control over, and true desert for, our actions ultimately rests on control over, and true desert for, our beliefs. A certain class of beliefs is especially relevant in this respect, namely evaluative beliefs. Evaluative beliefs are beliefs with an evaluative content. This evaluative content should include an agent's conception of a human life that is worth living and have potential effects as a criterion for choice and a guide for action. To insist, a crucial advantage of grounding moral responsibility on evaluative beliefs, rather than choices, is that the problem of an infinite regress of choices, as well as the correlative problem of a groundless, arbitrary choice as a basis for moral responsibility, do not need to arise. Before moving on to substantiate the proposal, however, let me point out that it is not without precedents. A step in this direction is Gary Watson's emphasis on values instead of – even second-order – desires in his view of moral responsibility (cf. Watson 1982), as well as Susan Wolf's 'Reason View' (the term is hers), according to which the ability to form correct values is necessary for moral responsibility (cf. Wolf 1990: 75), among others.

Let me now proceed to a defence of my proposal. If evaluative beliefs are to ground the possibility of ultimate control over our choices and actions, they have to satisfy a number of conditions, which would seem to include at least the following: 1) Corresponding to the depth of moral responsibility attributions, they should be a deep, core component of a person's self. 2) Under certain circumstances, the agent could be correctly considered as their true author and source. 3) The agent should have rational control over these beliefs. 4) The preceding condition should hold even if the beliefs are not causally determined (the proposal should not fall prey to some version or other of the so-called *Mind* argument).

I would like to argue that evaluative beliefs are able to satisfy these conditions. Unfortunately, owing to space limits, I can only give some brief remarks in favour of this contention.

Evaluative beliefs would certainly seem to satisfy the first condition. Our evaluative views are a central core of what we are, mentally speaking. No psychological

characterization of a person could be minimally complete unless it included a description of what she finds worth pursuing or avoiding in life. And if we reflect on our serious ascriptions of moral responsibility for choices or actions, we shall find ourselves ultimately praising or blaming an agent on account of the evaluative views these choices or actions express. In fact, if we come to see an act of hers as a momentary, passing impulse, not expressing her deep evaluative convictions, our judgement is significantly softened, or even withheld. This takes us to the second condition, which corresponds to the 'ultimate source' or true authorship aspect of ultimate control. In holding an agent responsible for an action on account of her evaluative views, we certainly seem to assume the she has proper control over them, so that she can be truly considered as her author and source. If we take this assumption to be false, we modify or even withhold our moral responsibility ascription. This happens in CNC manipulation cases, but also in more ordinary situations in which we do not see an agent as truly responsible for her evaluative beliefs. Certain victims of severely deprived childhoods or of a fanatical education can be examples of this predicament. But what requirements should be met in order for an agent to have deep control over, and to be the true author of, her evaluative views? In forming our initial evaluative views we are deeply influenced by our parents, close relatives or friends. This 'external', social origin is not, as such, a reason to deny our authorship with regard to the evaluative beliefs we end up having, as the student and Descartes examples show. But something else is needed. If we could do nothing but have the views we receive from our social environment, we would not be truly responsible for having them, and moral responsibility for our actions would not be possible. But we do not only receive beliefs, either evaluative or merely factual. We also acquire general standards that guide us in forming, assessing, accepting and rejecting such beliefs. This is an essential contribution that sociality makes to our constitution as agents. The most basic of these standards has to do with the truth-aiming character of belief. It demands of us that we allow our beliefs to be determined and controlled by the way things actually are. Another important standard for accepting and rejecting beliefs has to do with their logical relations. Contradictions, for example, are important reasons to modify our beliefs in order to avoid them. Applying these and other standards an agent can arrive to a system of beliefs which can truly be said to be her own and which she can be truly responsible for. And this holds for her evaluative beliefs as well. She may discover by her own, often painful, experience that an evaluative view she was taught is not actually true. And she may notice contradictions in her received evaluative views and be led to reshape them.

However, as our previous examples indicate, it is not needed, for these beliefs to be truly attributable to an agent, that she has chosen the standards and procedures she employs in forming, retaining or abandoning them. In a modest way, as compared with such great intellectual accomplishments as Descartes' *Meditations*, but not in a completely different sense, a system of evaluative beliefs can be truly ascribed to an agent as its source and author, and so as something for which she can truly deserve praise or blame. Moreover, though I shall not develop this important point here, I hold that having available alternatives to her actual evaluative beliefs is also a necessary condition on an agent's true authorship and responsibility for them. An alternative possibilities condition is applicable to an agent's evaluative beliefs no less than to her actions and choices if she is to have ultimate control over them.

The third condition corresponds to the other aspect of ultimate control, namely rational control. Irrational or arbitrary evaluative beliefs would not be fit to ground moral responsibility. Now, we control our actions rationally and voluntarily by choosing them in the light of our evaluative beliefs, but the control we should have over these beliefs in order for us to have ultimate control over our actions does not rest on a further choice of these beliefs. It is rather a matter of sensitivity to their internal consistency and respect for the facts they aim to capture, namely facts about what is valuable and worthwhile in human life. This control does not relate to our ability to choose and act, but rather to our ability to *see* what is there to be seen. This sort of non-voluntary, *theoretical* control, as it might be called, is an appropriate basis for praise and blame. It may relate to some cases in which we are blamed (or praised) for something we involuntarily did or omitted, such as forgetting (or remembering) our partner's birthday or an appointment we had for dinner. Even if believing, like forgetting, is not voluntary, and not even an action, we can justifiably blame someone for her beliefs, in the same way as our partner can justifiably blame us for forgetting our appointment. This blaming seems to assume some form of non-voluntary control. In blaming us for forgetting our appointment, our partner is blaming us for not *seeing* rather than not acting: for our blindness to both our appointment and her, which reveals our lack of consideration and respect for both. And she is assuming that we ought, and could, have remembered the appointment. Something closely analogous, I would think, holds when we blame someone for the evaluative views that an act of hers reveals.

In commenting on the last two conditions we have crucially insisted, in analogy with the case of our theoretical beliefs, on a sort of control based, not on choices or acts of will, but rather on an attitude of humility and respect to what is there, to something endowed with a kind of objectivity. Defending this sort of control seems to commit us to some version of objectivism, to the idea that there are facts of the matter that our beliefs should be guided by and respond to. In the case of evaluative beliefs, we are committed to some version of objectivism about the evaluative. An unrestricted, rampant objectivism in this field might appear as a rather implausible position. I do not think, however, that the version of objectivism we are committed to by virtue of our proposal should go that far. We do not need to assert that there are such entities as values. It seems enough to accept that there are facts of the matter in virtue of which evaluative beliefs can be true or false; that someone can be genuinely wrong in her evaluative views; and that evaluative truths are discovered, not invented.[4] However, it is important to note that someone can have the sort of control over her evaluative beliefs that is required for true authorship and desert even if these beliefs are not actually true. Essential to the possession of this sort of control is rather the attitude of respect to what is objectively there, as well as the aim of having one's beliefs guided and determined by it. This holds also for theoretical beliefs. To come back to a previous example, consider that many theses contained in Descartes' *Meditations* might actually be false, but, even if they were, Descartes would still be rightly considered as the true author of this great intellectual legacy and as truly praiseworthy for it.

A second issue raised by my proposal has to do with motivation.[5] If evaluative beliefs are to make an agent's ultimate control over her actions possible, they have to be able to motivate her to perform those actions. A complete causal-explanatory isolation between evaluative beliefs and actions would deprive an agent from control over the latter by means of the former. Motivation, like objectivity, is a big issue, which I cannot deal with in depth here and much less provide something resembling a solution to it. I will restrict myself to a couple of remarks. First, there is something quite bizarre in holding an evaluative belief with no motivating consequence at all, even potential. If, for example, someone seriously holds that causing unnecessary pain to an innocent person is morally wrong but does not feel motivated at all to act (and react) according to this (true) belief, there is reason to resort to psychopathology in order to account for this mismatch. She may feel motivated to act against such a belief (owing maybe to a sadistic tendency) and

actually do so, but if she really has the belief, only psychopathology could be of help if its contrary motivating potential did not manifest itself at least as shame or remorse. Second, evaluative facts need the participation of emotions in order to be properly appreciated. So, someone unable to feel or react emotionally in seeing someone else causing unnecessary pain to an innocent person is likely to be unable to see the wrongness in this situation and to form the corresponding general evaluative belief. Appreciation of evaluative facts and forming evaluative beliefs is a 'thick', emotionally laden cognitive performance. But the motivating potential of emotions is generally acknowledged.

Before moving on to the next condition, let me address another objection that might be raised against my proposal. I have insisted on a form of control over our beliefs based on respect and deference towards objective facts and principles. On this view, control would not consist in choosing these factors, but rather in being guided by them. It might be objected, then, that this view leads rather to heteronomy, to the subject's being 'remotely controlled', ruled by external forces.[6] I would like to respond as follows. First, it is important to point out that the case for my proposal rests in a large degree on the intuitions raised by the examples we considered. Now, are we really prepared to hold that, in following correct reasoning principles, mathematical rules, in accepting certain data and laws of physics as given in order to find out the result of the problem she was trying to solve, our student acted in a heteronomous way or that she was 'remotely controlled' by those objective, 'external' factors? My response, which I hope will be widely shared, would be 'no'. And a similar question could be raised and a similar answer given for what regards the example of Descartes. Second, I tend to think that the objection takes for granted the will-centred and individualistic conception of control that I have tried to dispute and so begs the question against my proposal. Someone who finds this objection powerful is probably in the grip of the individualistic and voluntaristic assumptions that I have been at pains to undermine, according to which an autonomous and morally responsible agent has to be a radically self-made entity, so that the influence of any factors beyond the scope of her choice is potentially threatening to her authorship and true desert for her deeds and accomplishments. Many factors coming 'from outside' are not in the way of our nature as morally responsible agents, but are rather constitutive of it. Again, if someone disagrees, I would ask him to reflect on the examples we presented and other similar cases and to judge whether many 'objective' factors present in the student's task of solving the problem or in Descartes' writing of the *Meditations* detract from their

respective praiseworthiness for those achievements or are rather constitutive conditions thereof. I would certainly say the latter, and I think there are powerful considerations in favour of this judgement. However, as I already pointed out, I do not have, at present, any conclusive arguments to offer to someone who rejects the judgement.

Let me finally address the fourth condition I mentioned. A traditional compatibilist objection to libertarianism is that indeterminism erodes an agent's rational and volitional control over her choices and actions, which turn into arbitrary, hazardous events, thus undermining her moral responsibility for them. This objection, which is usually known as the *Mind* argument, has been formulated in several ways. Consider Josephine, a judge who, after careful and relevant deliberation, decides at a certain time, T, not to grant clemency to a convict and to sentence him to life prison.[7] Let me now apply to this example Alfred Mele's recent version of the objection. Imagine a close possible world, with the same past and natural laws as the one Josephine inhabits, in which an identical twin of hers, call her Josephine*, exists. The first difference between the two worlds occurs only at time T. At this moment, while Josephine decides to sentence the convict to life prison, Josephine* decides instead to grant him clemency. On the assumption that Josephine's decision is causally undetermined, this is clearly conceivable. But then, in Mele's words, 'if … there is nothing about the agents' powers, capacities, states of mind, moral character, and the like that explains this difference in outcome, then the difference really is just a matter of luck' (Mele 1999: 99). From this perspective, Josephine's decision would appear to be a chancy and arbitrary event, not under her rational and volitional control.

This objection may be powerful against will-centred views of ultimate control, but the cognitive approach I recommend might succeed against it. On this approach, it is practical judgement, rather than choice, that plays the pivotal role in practical deliberation. A practical judgement about which action is best, or better than the alternatives, should be understood as the application of an agent's evaluative beliefs, as normative standards, to the situation she faces. Now, even if choices are not causally determined by practical judgements, it is not a mere accident that an agent's choice usually accords with her practical judgement. Practical judgement is a normative standard for choice and this is why discrepancy between the two usually counts as irrational and abnormal, as a case of incontinence or weakness of the will. Now with these remarks in mind we are not forced

to conclude that Josephine's decision was a matter of luck just because it differed from Josephine*'s. Since the only difference between the two worlds arises only at T, the decision's time, we can assume that Josephine and Josephine*, after a similar careful and relevant deliberation, formed the same practical judgement, namely that sentencing the convict to life prison was better than the alternatives. If, thereafter, Josephine decided in accord with this judgement while Josephine* decided against it, this only shows that Josephine*'s decision was irrational or weak-willed, but not that Josephine's decision was so as well. Josephine, unlike Josephine*, had rational control over her decision, which so was not arbitrary or merely lucky. The example may show that Josephine, like any of us, is not immune to irrationality or weakness of the will, but this is much less than is needed to show that indeterminism is incompatible with rational and volitional control over our choices and actions. What needs to be shown is that indeterminism turns virtually all of our choices and actions into lucky, arbitrary events. And the example does not yield this result.

I conclude, then, that a cognitive approach to moral responsibility, within the lines we have drawn, may be in a better position than a conative, will-centred approach in order to overcome some traditional difficulties of libertarianism, especially those related to ultimate control as a requirement for moral responsibility.

NOTES

[*] This paper is inspired in Moya (2006). Chapter 5 of this book is especially relevant for the ideas I present here. I thank Dr Lumer and Dr Nannini for their kind invitation to take part in the conference "Intentionality, Deliberation and Autonomy. The Action Theoretic Foundation of Practical Philosophy", in which I presented the original version of this paper. I thank Michael Bratman, Alfred Mele and the participants in the conference for their comments, suggestions and criticisms. Very special thanks are due to Christoph Lumer for reading carefully the original version of this paper and suggesting many ways in which it could be improved.

[1] Though the point would need more argument, it may suffice to note that control, so understood, would be possessed by higher animals and very young children.

[2] For purposes of exposition, I shall present Kane's work as if it were an attempt to respond to Strawson's argument. Of course it is much more than that.

[3] Bernard Williams (1973) famously argued that beliefs are not under our direct voluntary control, so that there is no much room for deciding to believe. I find Williams's rejection of doxastic voluntarism very convincing.

[4] Objectivism about evaluative beliefs or judgements is far from being a hopeless position. I would tend to think that it is rather subjectivism, in its different forms, that is really in trouble. A recent and solid defence of objetivism (and cognitivism) in ethics can be found in Wiggins (2005).

[5] Christoph Lumer encouraged me to discuss this and the previous issue concerning objectivism.

[6] It was also Christoph Lumer who raised this objection. The terms 'heteronomy' and 'remotely controlled' are his.

[7] The example is vaguely inspired in an example of Van Inwagen's (1983: 68–9).

REFERENCES

Clarke, R. (1997) 'On the possibility of rational free action'. *Philosophical Studies*, 88: 37-57.

Gilson, E. (1951) *Etudes sur le rôle de la pensée médiévale dans la formation du système cartésien*. Paris: Vrin.

Kane, R. (1996) *The Significance of Free Will*. Oxford and New York: Oxford University Press.

Mele, A. R. (1999) 'Kane, luck, and the significance of free will'. *Philosophical Explorations*, 2: 96–104.

Moya, C. J. (2006) *Moral Responsibility: The Ways of Scepticism*. Abingdon: Routledge.

Pereboom, D. (2001) *Living Without Free Will*. Cambridge: Cambridge University Press.

Strawson, G. (1986) *Freedom and Belief*. Oxford: Clarendon Press.

Van Inwagen, P. (1983) *An Essays on Free Will*. Oxford: Clarendon Press.

Watson, G. (1982) 'Free agency', in G. Watson (ed.), *Free Will*. Oxford: Oxford University Press.

Wiggins, D. (2005) 'Objectivity in ethics; two difficulties, two responses'. *Ratio (new series)*, 18: 1–26.

Williams, B. (1973) 'Deciding to believe', in his *Problems of the Self*. Cambridge: Cambridge University Press.

Wolf, S. (1990) *Freedom Within Reason*. Oxford: Oxford University Press.